ENTROPY-WEIGHTED RANDOM FORESTS: ROBUST LEARNING IN NOISY, IMBALANCED DATA

Anusha Agarwal, Viola Pande

OUTLINE

01

Introduction

- Individual decision trees overfit
- Random Forests combat this by aggregating predictions
  - High-dimensional, noisy data → failure
  - Biological, chemical, medical fields
- What if we weight votes based on **entropy** rather than **majority**?
  - Input: chemical compound
  - Output: active / inactive (binds to thrombin or not)
  - Working **through** the noise rather than getting rid of it

02 Related Work

1. Performance on similar data → 2% accuracy increase (Evanthia et. al, 2010)

2. General classification rate → 1-2% accuracy increase (Evanthia et. al, 2010 adapted from Šikonja, 2004)

3. F1-score → 1.25% accuracy increase (Raushan, 2017)

4. Accuracy and probability-based confidence intervals → 1.21% accuracy increase (Zhang et. al, 2023)

03 Dataset

- **DOROTHEA** - UCI Machine Learning Repository: chemical compound binary classification
- **Probe features**: Deliberate, irrelevant distractor features designed to introduce noise (100,000 features total, 1150 instances). 55/30/15 training, validation, testing split.
- **Binary features**: presence/absence of specifical chemical structures in each compound, no normalization or feature selection performed
- **High class imbalance**: 90% inactive compounds, 10% active compounds. Chosen to test EWRF on class imbalance challenges.

04 Methodology

01 Vanilla Random Forest

Built with scikit-learn, each decision tree is given EQUAL WEIGHTING when aggregating using majority voting scheme for final predictions.

02 Entropy-weighted RF

EWRF weights each tree based on prediction confidence. We calculate leaf-node entropy for each sample (Shannon) and apply the exponential formula below to inversely weight trees based on entropy

03 Probability Aggregation

Apply weights to individual leaf node distributions and aggregate the probability distributions across all classes, normalize, and choose class with highest weighted probability.

04 Experimentation

Tested with various maximum decision tree depths (5, 10, 15, 20, None) and number of trees (50, 100) to explore shallow + high bias alongside deep + high variance trees.

$$P(y = c \mid x) = \frac{\sum_{t=1}^{T} w_t(x) \bullet p_t(c \mid x)}{\sum_{t=1}^{T} w_t(x)}$$

$$H = -\sum_{i=1}^{k} p_i * \log_2(p_i)$$

$$w = exp(-\alpha H)$$

05 Results

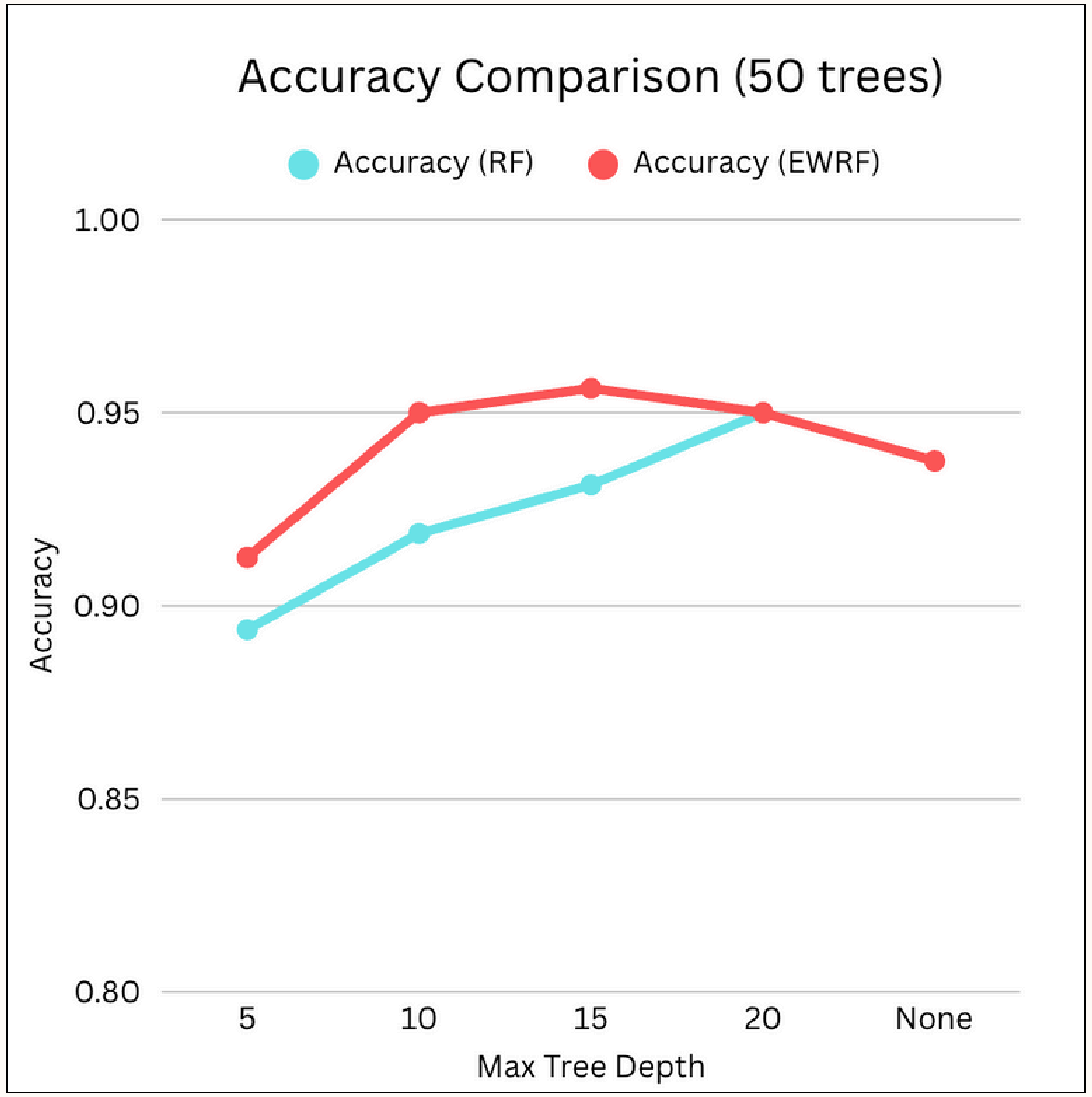
Accuracy: 96.25% Improvement (A): 4.38% Precision: 0.8125 Recall: 0.8125 F1-score: 0.8125 Improvement (F): 37.77%	$\begin{bmatrix} 141 & 3 \\ 3 & 13 \end{bmatrix}$

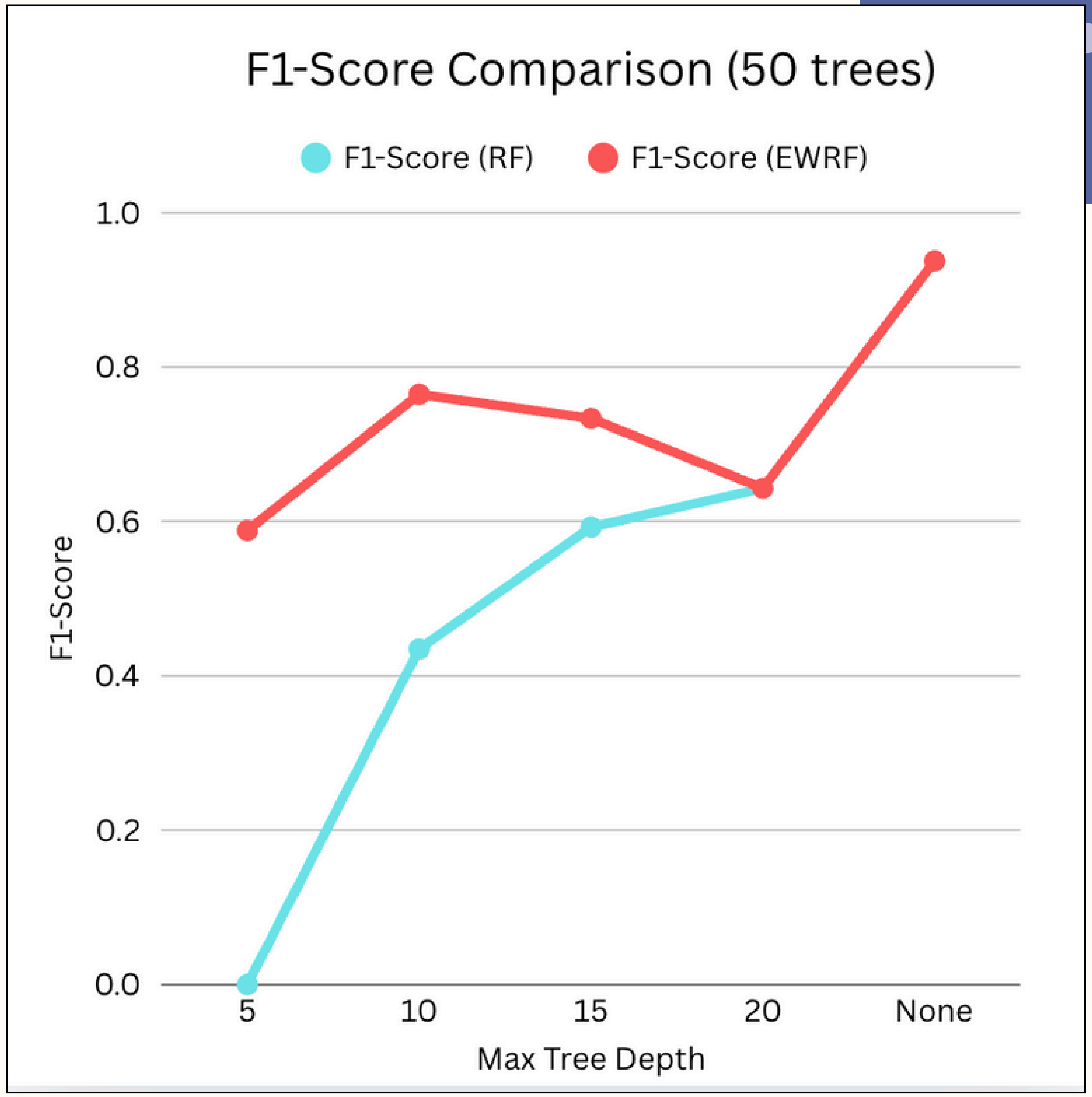
Best run: moderately depth-constrained (d=10)
This is when entropy-weighted voting has the greatest effect on accuracy!

Accuracy: 92.5% Improvement (A): 0% Precision: 0.7000 Recall: 0.4375 F1-score: 0.5385 Improvement (F): 0%	$\begin{bmatrix} 141 & 3 \\ 9 & 7 \end{bmatrix}$

Worst run: unlimited tree depth
Entropy becomes a useless metric

06 Results (cont.)





07

Discussion

01

Shallow depths

Highly uncertain predictions from RF, trees with high entropy leaf nodes dominate the prediction. Exponential down-voting leads to clear performance gain.

02

Overfitting

At tree depths >=15, model improvement is diminished. Trees are more prone to overfitting, leading to artificially low entropies. Our core assumption of low entropy = confident predictions breaks.

03

Class imbalance

Extremely high improvements in F1-score, vanilla RF has clear majority class bias. Accuracy vs F1-score improvement shows strong performance in tackling class imbalance issue in RF.

04

Study comparison

Highest observed increase in accuracy, 4.38%, was over double that found in any related work. F1-score more applicable to noisy datasets; results just as promising. Emphasizes balanced gains across both precision and recall.

08 Conclusions

STANDARD RF

Great for combatting overfitting issue with decision trees; falters on more complex datasets

ENTROPY-WEiGHTED RF

More accurate and precise on average, extending benefits of Random Forests to domains with noisy, high-dimensional data → medical and chemical applications

CAVEAT

Changes in accuracy across depth constraints solidifies our understanding of how entropy is most useful on moderately complex trees

09

Future Directions

- Study effects of weighted voting with different models of entropy, such as Tsallis or von Neumann
- Combine our developed EWRF formula with other metrics such as accuracy or F1-score
- Test multiple datasets → increase generalizability and reliability through replication, gaining a more comprehensive review across databases

THANK YOU!